WIDE AREA NETWORK MONITORING SYSTEM FOR HEP EXPERIMENTS AT FERMILAB

Maxim Grigoriev, FNAL, Batavia, IL 60510, USA

Les Cottrell, Connie Logg, SLAC, Menlo Park, CA 94025

Abstract

Large, distributed High Energy Physics (HEP) collaborations, such as D0, CDF and US-CMS, depend on stable and robust network paths between major world research centres. The evolving emphasis on data and compute Grids increases the reliance on network performance.

Fermilab's experimental groups and network support personnel identified a critical need for WAN monitoring to ensure the quality and efficient utilization of such network paths. This has led to the development of the Network Monitoring system we will present in this paper.

The system evolved from the IEPM-BW project, started at SLAC three years ago.

At Fermilab this system has developed into a fully functional infrastructure with bi-directional active network probes and path characterizations.

It is based on the Iperf achievable throughput tool, Ping and Synack to test ICMP/TCP connectivity. It uses Pipechar and Traceroute to test, compare and report hop-by-hop network path characterization. It also measures real file transfer performance by BBFTP and GridFTP. The Monitoring system has an extensive web-interface and all the data is available through standalone SOAP web services or by a MonaLISA client.

Also in this paper we will present a case study of network path asymmetry and abnormal performance between FNAL and SDSC, which was discovered and resolved by utilizing the Network Monitoring system.

NETWORKING AT FERMILAB

Overview

Fermilab [1] is the largest US laboratory for research in HEP. Every second hundreds of Megabytes of physics data are flying through FNAL's LAN to and from the world's biggest research labs such as CERN, BNL, ANL and SLAC as well as to hundreds of physics institutions. This is a truly distributed scientific environment where each person depends on the quality and robustness of current network paths. With the upcoming LHC era, future development of the DOE UltraScienceNet[2] and the need to bring HEP Grids to the desktop, the emphasis on availability and performance of computer networks is increasing every day. An example of the current utilization of Fermilab's networks is shown in Fig.1.

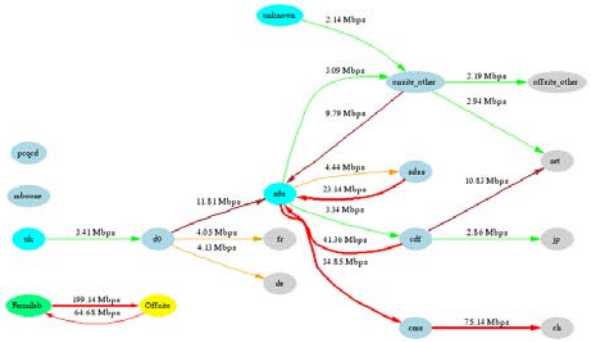

Figure 1: Snapshot of Fermilab's network traffic

WAN Topology, Experiments

The Fermilab network connects a wide variety of heterogeneous computer resources. It's a multi-subnet intranet with more than 9000 computers connected by an ESnet provided OC12 uplink to the world. This connection needs to be tuned and monitored 24 hours by 7 days/week.

The current data flows for D0 and CDF experiments

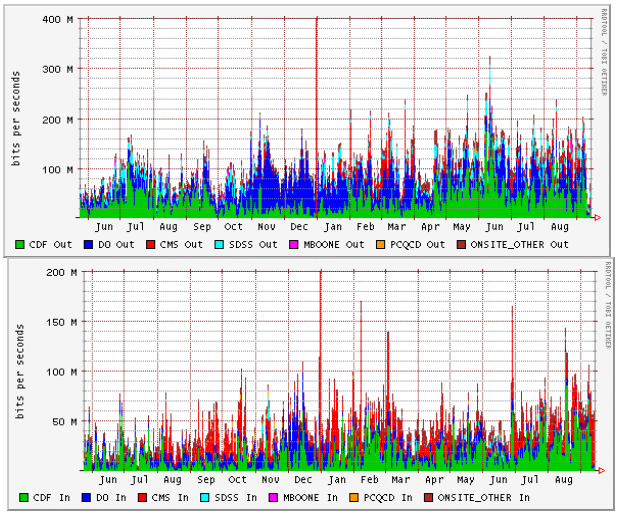

Figure 2: Outbound and Inbound traffic rates for the past year to and from Fermilab's HEP collaborations

are about 50Mbps inbound and 150Mbps outbound (see Fig.2). The major consumers of the data are in the US, UK, France, Italy, Germany and Canada. We anticipate an almost exponential increase in the average data rates every year. Also there is an increasing impact on network performance from the CMS experiment. The CMS experiment has upcoming data challenges for robust transfer rates starting from 100 Mb/sec.

MONITORING SYSTEM

Description

The Internet End-to-End Performance Monitoring Bandwidth to the World (IEPM-BW)[4] project started at SLAC about 3 years ago.

The original purpose of the IEPM-BW [5] project was to develop and use an infrastructure to make active end-to-end application and network performance measurements for high performance networks such as those used worldwide by HENP and Grid applications.

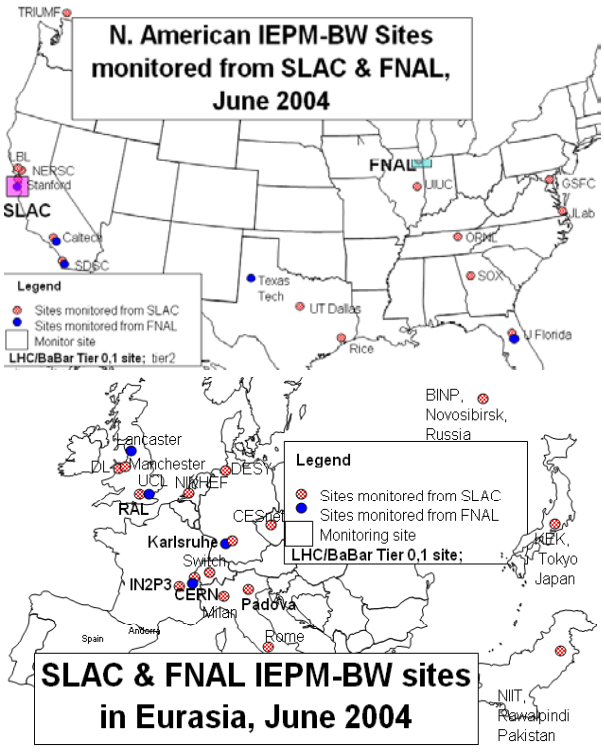

Figure 3: IEPM-BW monitoring hosts

The specific goals were to:

- Develop a simple, robust, ssh based infrastructure providing regular ongoing measurements from a selection of tools such as iperf, traceroute, BBFTP etc.;
- Develop data reduction, analysis, reporting, forecasting, archiving, visualization and publishing tools;
- Evaluate measurement tools to determine their applicability, network impact, accuracy and ease of use;
- Allow customization at each monitoring site to select the tools to be run, the scheduling frequency, and the remote hosts to be monitored.
- Provide information in a form suitable for: trouble-shooting, understanding network performance, planning and setting expectations.

Fig. 3 shows the current sites in 13 countries that are monitored from SLAC and FNAL. Besides SLAC and FNAL, IEPM-BW has been installed and measurements made from eight other sites. The total number of host-to-host pairs for SLAC and Fermilab is about 100.

After extensive use and evaluation IEPM-BW was modified to better fit Fermilab's needs. Network probes, based on BBCP file transfers were removed while tests based on GridFTP file transfers and Pipechar active monitoring were added. All tests were extended to run in both directions. The Web visualization interface was customized and on-the-fly generation of the performance history graphs was added. Currently, at Fermilab, there are 6 different active network performance monitoring bi-directional tools. Current performance monitors includes:

- ***Ping*** (or SYNACK [6] in cases where ping is blocked), to get RTT/Loss rate and general connectivity (if it fails then no further tests are done for the node)
- ***Traceroute*** – to obtain a list of the hops in both directions, and RTTs to each hop(with hop-by-hop analysis of the results)
- ***Pipechar*** - to obtain a list of hops in both directions, and RTTs to each of hop with Bandwidth on every hop, utilization rate and bottleneck analysis [7]
- ***Iperf*** – to obtain achievable bandwidth to and from the monitored node [8]
- ***BBFTP*** disk-to-disk bi-directional file transfers (with files ranging in size from 100MB to 500MB, depending on available bandwidth information), to determine the real file transfer performance [9]
- ***GridFTP*** disk-to-disk bi-directional file transfers [10]

Visualisation

For an end-user, network performance information is presented in many forms. The most useful are time series plots and tables of the last obtained results. Also, there are scatterplots of each result set to each metric compared to another (e.g. Iperf vs. RTT,RTT vs.

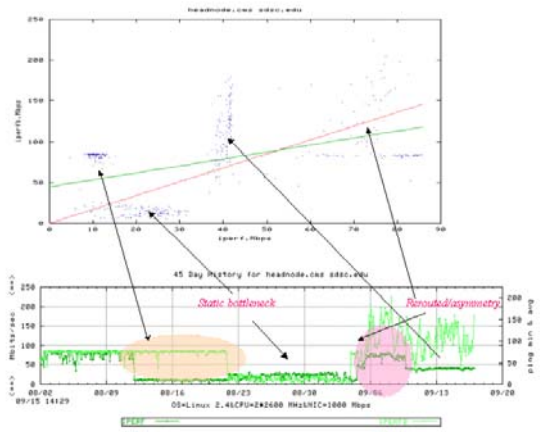

Figure 4: Direct and reverse Iperf test results presentation as scatter plot and time series

BBFTP etc) to show any correlations between them (see Fig.4.).

Each results set is also presented as a histogram to show the most expected performance value of the network probe. There are bi-directional traceroute and pipechar analysis tables, shown to alert a user of any sudden network path changes and path asymmetry (if the number of hops changed or any of the per-hop values changed by more than 30%). The whole monitoring website is logically divided up by HEP experiments. New hosts to be monitored are easily added by adding a few lines into the configuration file.

Analysis, Forecasting

It is very important to have a monitoring system with adequate response time to network anomalies. Because of the nature of the system (2 hour monitoring breaks) we didn't set the goal of the identifying any kind of short-lived network problems, instead the whole idea behind the results analysis is to determine significant and consistent (i.e. long term changes, persisting for several hours at least) shifts in network performance. After researching through numerous publications on statistical network analysis and forecasting techniques [13], [15], [16], the tri-exponential approximation with moving time frame of the results with special rules set was developed. It is based on combining the forecasting technique, employing the triple-exponential smoothing as described by [13], [14] with the X^2 error estimation method, and applied to normalized Iperf test results.

The Iperf test results were chosen as the most accurate indicator of the achievable throughput. The X^2 criteria is applied to the moving window of the last 7 observations (14 hours). See Fig.6 for a graph with normalized Iperf, upper and lower limits, outstanding observations, X^2 graph and alerts marker.

Setting Alerts

To set up an alert a set of rules and procedures was developed. First of all, the system checks for the lost tests and reports them, then for every new data point the set of forecasted values for $N_{total} - N_{time\text{-}window}$ is built and for the last $N_{time\text{-}window}$ values the X^2 sum is calculated. The alert is generated if X^2 sums are more than some threshold (set to cut off small variations). If there is $N_{time\text{-}window} - 1$ consecutive alert, then the system sends a message to the sysadmin, notifying about a significant drop in the network performance.

Availability of results

All results are available on the website [3]. Also, all data is available through Webservice requests, implemented as SOAP [12] over an HTTP server. The location of the Webservices access is http://dmzmon0.deemz.net/~wanbanmon/soap/wsdl/IEPM_profile.wsdl and the schema is compliant with http://www-didc.lbl.gov/NMWG/docs/GFD-R.023.pdf. Requests can be sent for the parameters:

- ***path.bandwidth.achievable.TCP*** (Iperf, reverse Iperf),
- ***path.bandwidth.achievable.TCP.multistream*** (BBFTP and GridFTP bi-directional)
- ***path.bandwidth.capacity*** (pipechar).

More on IEPM-BW webservices can be seen at http://www-iepm.slac.stanford.edu/tools/web_services. In addition, all monitoring statistics also presented by the MonaLISA [11] agent, see Fig.5.

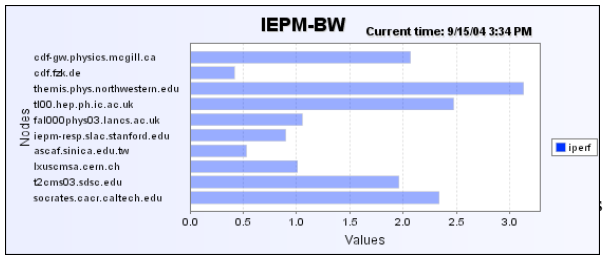

Figure 5: View of the IEPM-BW by MonaLisa client

CASE STUDY

Problems with link to SDSC

The network path from FNAL to SDSC had several problems in the past. The most significant problem was path asymmetry. This appeared from time to time due different routing by ESnet and therefore asymmetry in the throughput caused by a bottleneck between Esnet and CENIC. Also monitoring suffered from the limitations of the remote node itself (TCP settings were far from optimal).

All changes are identified and shown in Fig. 6.

System Response, resolution

Monitoring system response is clearly seen on the Fig.7. The red spikes correspond to actual alert conditions raised from loss of the tests (host unreachable due the firewall) or drastically dropped performance due the monitored host limitations.

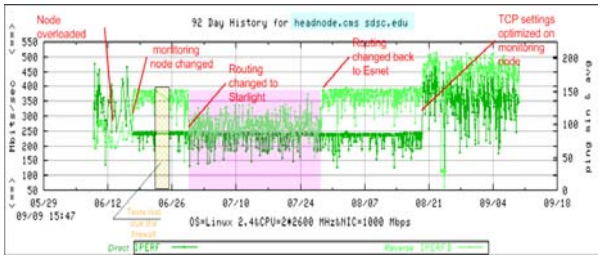

Figure 6: Direct and Reverse Iperf test results for SDSC link

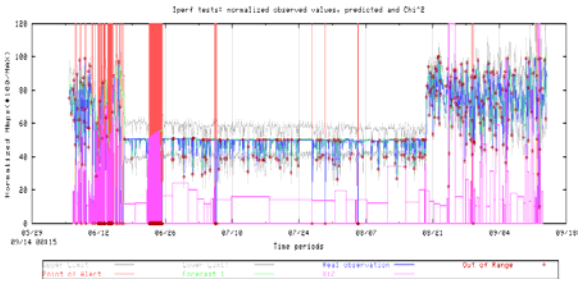

Figure 7: Alerts plot and messages for SDSC link

The following table (Table 1) shows alert messages received by sysadmin.

Table 1

Dropped performance Alert:	Please take a look on headnode.cms_sdsc.edu iperf.Mbps graph at- 06/22/2004 12:44:01 the throughput is dropped to 50.8 Mbps
Lost test Alert:	Please take a look on headnode.cms_sdsc.edu iperf.Mbps graph at -06/17/2004 09:23:33 some tests (for more than 5 hours) were lost
Dropped performance Alert:	Please take a look on headnode.cms_sdsc.edu iperf.Mbps graph at - 06/16/2004 12:44:01 the throughput is dropped to 97.4 Mbps

FUTURE PLANS

We look forward to reducing of the amount of monitoring traffic. One of the possible options is substitution of the highly intrusive Iperf network probe with the ABwE [17] bandwidth estimation tool. Also, traceroute analysis needs to be improved. Corresponding work by Les Cottrell and Connie Logg [18] looks promising and more visually advanced.

The WAN monitoring system was designed to be helpful tool for every participant of HEP. If you are interested in setting up a monitored node at your site, please submit your requests to mailto:iepm-bw@fnal.gov.

ACKNOWLEDGEMENTS

We would like to thank Phil Demar and Donna Lamore for their comments and valuable additions on the nature of Fermilab's networking. Special thanks to Jerrod Williams of SLAC and to all sysadmins of the remote monitoring sites for their support.

This work was supported by U.S. DOE Contract no. DE-AC02-76SF00515.

REFERENCES

[1]http://www-dcn.fnal.gov/.

[2]http://www.csm.ornl.gov/ultranet.

[3]Fermilab's WAN monitoring webpage[4] IEPM-BW, http://www-iepm.slac.stanford.edu.

[5] R. Les. Cottrell and Connie Log, "Overview of the IEPM-BW Bandwidth Testing of Bulk Data Transfer", SLAC –PUB-9202, SLAC, July 2003

[6] SYNACK

[7] Pipechar, http://www-didc.lbl.gov/NCS.

[8] Iperf, http://dast.nlanr.net/Projects/Iperf.

[9] BBFTP, http://doc.in2p3.fr/bbftp.

[10] Globus toolkit, http://www.globus.org.

[11] H.B. Newman, I.C. Legrand, P. Galvez, R. Voicu, C. Cirstoiu, " MonALISA: A Distributed Monitoring Service Architecture", Proceedings of CHEP 2003, La Jolla, Ca, USA, March 2003.

[12] perl module SOAP::Lite

[13] Jake D. Brutlag, "Aberrant Behaviour Detection in Time Series for Network Monitoring", Proceedings of LISA 2000, New Orleans, LA, USA, December 2000.

[14] NIST e-handbook of statistics, http://www.itl.nist.gov/div898/handbook.

[15] McGregor A.J. and Braun H-W, "Automated Event Detection for Active Measurement Systems", Proceedings of PAM2001, Amsterdam, Netherlands, April 2001

[16] P. Barford, J. Kline, D. Plonka and A. Ron, "A Signal Analysis of Network Traffic Anomalies", Proceedings of the second ACM SIGCOMM Workshop on Internet measurement, Marseille, France, 2002

[17] Jiri Navratil and R. Les. Cottrell, "ABwE: A Practical Approach to Available Bandwidth Estimation", Proceedings of PAM2003, San Diego, USA, April 2003

[18] Traceanal: a tool for analyzing and representing traceroutes presented by Les Cottrell at the Internet 2 Joint Techs E2Epi BOF Columbus Ohio, July 2004.